DOCUMENT RESUME

ED 421 498 TM 028 457

AUTHOR Price, Larry R.; Oshima, T. C.
TITLE Differential Item Functioning and Language Translation: A Cross-National Study with a Test Developed for Certification.
PUB DATE 1998-04-00
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Certification; Cross Cultural Studies; *Cultural Differences; Diving; International Studies; *Item Bias; Tables (Data); Test Construction; *Test Format; Test Items; *Translation
IDENTIFIERS Item Bias Detection; *Japanese People

ABSTRACT

Often, educational and psychological measurement instruments must be translated from one language to another when they are administered to different cultural groups. The translation process often necessarily introduces measurement inequivalence. Therefore, an examination may be said to exhibit differential functioning if the test provides a consistent advantage to one particular race or culture through the manner in which the test items are written. One thousand American and 1,134 Japanese entry-level examinees participating in a scuba diving certification course took a standardized criterion mastery test for certification. The parametric framework Differential Functioning of Items and Tests (DFIT) proposed by N. Raju, W. van der Linden, and P. Fleer (1992) was used to detect differential item functioning (DIF). Out of a total of 30 items, 10 were found to exhibit significant noncompensatory DIF. Differential test functioning was also found to be significant. This paper demonstrates that the new DFIT technique can be applied successfully to the translated data, and that possible causes for the differential functioning can be examined using results from the DFIT analysis. (Contains 3 figures, 5 tables, and 25 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 498

Running Head: DIFFERENTIAL ITEM FUNCTIONING

Differential Item Functioning and Language Translation:

A Cross-National Study With a Test Developed for Certification

Larry R. Price

Emory University

T. C. Oshima

Georgia State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Larry Price
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper Presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA.

TM028457

Abstract

Often, educational and psychological measurement instruments must be translated from one language to another when they are administered to different cultural groups. The translation process often necessarily introduces measurement inequivalence. Therefore, an examination may be said to exhibit differential functioning if the test provides a consistent advantage to one particular race or culture through the manner in which the test items are written. One-thousand America and one-thousand one hundred thirty four Japanese entry level examinees participating in a scuba diving certification course took a standardized criterion mastery test for certification. Differential Functioning of Items and Tests (DFIT) proposed by Raju, van der Linden, and Fleer (1992) was used to detect Differential Item Functioning (DIF). Out of a total of thirty items, ten were found to exhibit significant Non-Compensatory DIF. Differential Test Functioning (DTF) was also found to be significant. This paper demonstrated that the new DFIT technique can be successfully applied to the translated data and that possible causes for the differential functioning can be examined using the results from the DFIT analysis.

Differential Item Functioning and Language Translation: A Cross-National Study With A Test Developed For Certification

Often, educational and psychological measurement instruments must be translated from one language to another when they are administered to different cultural groups. This translation process, necessarily introduces the problem of measurement equivalence. The application of item response theory (IRT) in the analysis of a translated test provides an opportunity for cross-cultural testers to solve the problem of measurement equivalence while simultaneously revealing important differences due to culture, language or a combination of both (Hambleton & Swaminathan, 1985; Lord, 1980). Furthermore, a test may be said to be culturally unfair if it provides a consistent advantage to one particular race or culture through the manner in which the test items are written.

Drasgow (1984) states that "measurement equivalence exists when the relations between observed test scores and the latent trait or attribute measured by the test are identical across subpopulations" (p. 134). In the case of translated tests, when individuals who are equal in the trait measured by the test, but who come from different cultural and linguistic groups have the same observed score, test are said to exhibit measurement equivalence. If measurement inequivalence is found, the test should be revised by improving or replacing inadequate items. Finally, only after the measurement equivalence is established, can differences between groups be examined.

Item Response Theory and the Detection of Differential Item Functioning

Item response theory has been applied to a variety of translated tests in order to evaluate measurement equivalence (Candell & Hulin, 1987; Ellis, 1989; Hulin, 1987; Hulin, Drasgow & Komocar, 1982; Hulin & Mayer, 1986, Budgell, Raju & Quartetti, 1995). The term that has evolved from item response theory literature that is used to describe the differential performance between groups at the item level is known as differential item functioning (DIF). In past studies, after DIF items have been identified, they are either corrected to eliminate DIF and returned to the item pool or eliminated entirely. Theoretically, the end result is a test that exhibits measurement equivalence. Finally, after DIF is detected within a test, judgmental and empirical evaluation must be conducted in order to explain the possible cause of DIF. Such explanation is crucial in order to help determine whether test scores represent group differences or measurement artifacts.

In IRT, the probability of a correct response on an item for an individual with a latent trait θ is described by an item characteristic function (ICF), the S-shaped curve. Additionally, the curve is usually defined by three parameters and represented by the logistic function:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (1)$$

where,

$P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly,

b_i is the item difficulty parameter,

a_i is the item discrimination parameter,

c_i is the pseudo-chance-level parameter,

D is a scaling factor designed to make the logistic function closely approximate the normal ogive function (D=1.702).

Figure 1 illustrates an example of a three-parameter model ICF with a = .425, b = 0, and c = .20.

Insert Figure 1 about here

The cornerstone of IRT rests on the property of invariance. This property implies that the ICF is unique under the conditions of a particular model except for random variations. Linn (1981) and Shepard (1987) agree that when the curves from the two groups differ the fundamental assumptions of item response theory models are violated. The assumptions of the item response theory model include unidimensionality (e.g., that the test measures only one construct). Evidence of large DIF indicates that an item is measuring an additional construct in one of the two groups, and the construct may not be relevant to the intended purpose of the test.

The aim of this study was twofold. The first purpose of this study was to empirically investigate the usefulness of the DFIT framework when comparing the original and translated versions of an internationally standardized scuba diving certification test. The second purpose of this study was to investigate the possible cause/explanation for DIF items on the scuba diving certification test. An English and Japanese version of an internationally used criterion-referenced mastery test for certification was the instrument of choice for this study. Differential item and test functioning based on a three-parameter IRT model using the entire ability range was the statistical evaluation technique employed (Raju, van der Linden & Fleer, 1992). Judgmental evaluation of differential functioning was addressed through: (1) the pattern of significant non-compensatory DIF (NC-DIF) items; and (2) content analysis of those items.

Differential Functioning of Items and Tests

Raju, et al.(1992) proposed a parametric framework, known as differential item functioning of items and tests (DFIT) that allows for individual item DIFs to add up to total test differential test functioning (DTF). Because the test item is the most fundamental part of a test, DIF studies are important in uncovering possible unfairness in test use. Furthermore, it is possible for several items in a test to exhibit DIF but the test to be fair. Therefore, potential unfairness at the test level (DTF) should also be examined. The following section describes the DFIT framework briefly. See Raju et al. (1992) for details.

Measures and Terminology of Differential Functioning of Items and Tests (DFIT)

Differential Test Functioning

Let $P_i(\theta_s)$ represent the probability of success for an examinee s with ability θ on item i (see equation 1). The test may consist of k items and have one set of item parameters for each of two groups (Reference Group and Focal Group). Further, an assumption is made that the two sets of item parameters are on a common scale. Let $P_{iR}(\theta_s)$ represent the probability of success on item i for examinee s as if he or she were a member of the Reference Group; likewise, let $P_{iF}(\theta_s)$ represent the same probability of success for the same examinee on the same item as if he or she were a member of the Focal Group. If an item is functioning differently in two groups, P_{iR} and P_{iF} should be different for some examinees.

Within IRT, an examinee's true score can be expressed as

$$T_s = \sum_{i=1}^{k} P_i(\theta_s) . \qquad (2)$$

Theoretically, in the present explanation, each examinee will have two true scores, one as a member of the Focal Group (T_{sF}) and the other as a member of the Reference Group (T_{sR}). If T_{sR} and T_{sF} are equal for an examinee, the examinee's true score is independent of membership. Furthermore, the greater the difference between T_{sR} and T_{sF}, the greater the

differential functioning of a test. A measure of DTF at the examinee level may be defined as $(T_{sF} - T_{sR})^2$. Therefore, an overall measure of DTF across examinees may be defined as

$$DTF = \underset{F}{\in} (T_{sF} - T_{sR})^2, \qquad (3)$$

where the expectation ($\in$) can be taken over the Focal Group.

Differential Item Functioning

The DFIT framework allows for the formulation of two measures, compensatory DIF (C-DIF) and non-compensatory DIF (NC-DIF). The two measures provide distinct but related types of information about the functioning of an item. C-DIF is related to the DTF as follows:

$$DTF = \sum_{i=1}^{k} C\text{-}DIF_i. \qquad (4)$$

Since DTF is the sum of C-DIF, there is a possibility for cancellation of differential functioning at the test level when one item displays C-DIF in favor of one group and another item displays C-DIF for the other group. In practical settings, a test developer can examine which C-DIF items need to be eliminated in order to reduce overall DTF.

NC-DIF on the other hand assumes all items other than the one under study are free from differential functioning. Therefore NC-DIF is not additive. In this sense, NC-DIF is

similar to other IRT-based DIF indicies such as Lord's chi-square (Lord, 1980) and area measures (Raju, 1988). Raju (1992) noted that items having significant NC-DIF do not necessarily have significant C-DIF. An example of this occurs when one item favors the Reference Group and another item favors the Focal Group. In this case, NC-DIF occurs, but C-DIF may not be significant.

DFIT Significance Test

Raju et al. (1995) proposed a significance test for DTF. A significant chi-square for DTF indicates that there exists a significant differential test functioning. When DTF is significant, one item (typically an item with large C-DIF) is removed and DTF is tested again. This process is repeated until the chi-square test shows no significance. Those items removed to achieve non-significant DTF are regarded as "significant" C-DIF items. Although a significance test for NC-DIF was theoretically described in Raju, et al. (1995), the authors recommend an empirical approach to declare the significance of NC-DIF (NC-DIF >.006) based on a simulation study by Fleer (1993).

Method

Item and Test Translation

The original item translation for the 50 item test was performed by Mr. Yoshinori Izumi, Training Manager for the National Association of Underwater Instructors (NAUI) Enterprises, Incorporated in Tokyo, Japan. Mr. Izumi was selected to translate the test from English to Japanese based on his qualifications as a content expert in the field of sport scuba diving along with his verbal and written fluency in both the English and Japanese

languages. Most of the items in the Japanese version were either precisely or loosely translated from items on the English version. However, some of the items were totally rewritten for the Japanese examinees. Since the DFIT analysis only makes since for those items with the same content but in the different language, Mr. Izumi identified 30 of the 50 items to be semantically and linguistically similar enough to be included in this study. The 30 item test included the following 6 content areas: (a) Skills/Safety, 13 items; (b) Decompression, 1 item; (c) Physics, 5 items; (d) Physiology, 6 items; (e) Equipment, 2 items; and (f) Environment, 3 items. Of those 30 items, 17 items had precise translation. Of the remaining 13 items, some items had stem and/or option differences while tapping on the same content area. Separate keys were used for the English and Japanese versions.

Participants

Data collection for this study was conducted during May through October, 1996 in Tokyo, Japan and in California, Georgia and Florida in the United States. The subjects participating in the study were Japanese and American males and females between the ages of 18 and 40 years enrolled in an entry level scuba diving certification test sanctioned by the National Association of Underwater Instructors (NAUI). The sample consisted of 1134 Japanese and 1000 American males and females. All subjects participating the study had a minimum of 12 years of formal education in their respective country's educational system. There were no students with mental or physical disabilities that participated in the study. This information was offered on the student's individual confidential course file that was completed prior to the course beginning.

Course Curriculum and Test Administration

Both groups received a course curriculum written by NAUI. The method of instruction used by the Instructors for conveying the information in the curriculum was approximately 70% lecture and 30% discussion. Visual teaching aids were used by the Instructors during all lectures and discussions.

The standard NAUI test for certification was administered to both groups at the end of the formal course of instruction. Subjects were given 1 hour and 30 minutes to complete the 50 item test. No notes or textbooks were allowed to be used as reference material during the test. Calculators were allowed in order to compute applied problems related to diving physics. Decompression tables were allowed to be used in order to complete the applied decompression problems on the test. All answers were recorded on a separate answer sheet with an identification number.

Data Analysis and Parameter Estimation.

After data collection was complete, test answer sheets were examined for errors and accuracy of answer coding. Next, parameter estimation and data analysis was conducted in the following sequential steps:

1. The Statistical Package for the Social Sciences personal computer software program (SPSS 7.5) was used to calculate descriptive statistics related to the demographics of the sample and to perform classical item analysis based on classical test theory. In addition, the reliability of the test was investigated.

2. Unidimensionality of the test was verified (χ^2 = .326) by the use of the DIMTEST (Stout, 1991) computer program.
3. BILOG 3.10-PC computer was used for the estimation of item and ability parameters. The program's Marginal Maximum Likelihood Estimation MMLE procedure was used for the estimation of item parameters under the three-parameter logistic model. Estimates of underlying ability were made via the program's Bayesian EAP procedure using the unit normal prior. BILOG goodness-of-fit indices were examined informally for model-data fit. Although the Japanese sample consisted of 1134 subjects, BILOG randomly selected 1000 subjects for the analysis.
4. After the item parameters were estimated, the Japanese and American examinees were placed on a common scale by the test characteristic curve method (Stocking & Lord, 1983) as incorporated into the computer program IPLINK (Lee & Oshima, 1996).
5. DIF and DFIT measures were computed using the framework proposed by Raju, et al. (1992).
6. DIF and DFIT indices were computed for theta ranges across the entire ability range. DIF and DFIT indices were computed using the estimated a-, b- and c-parameters. DIF measures computed included the chi-square statistic for DTF. For all measures, items were examined for significant differential functioning at the alpha level of .01. NC-DIF items were declared significant if they had a

value greater than .006. The .006 significance level was empirically established through a previous Monte Carlo study by Fleer, (1993).

Results

The results of this study are organized into four sections. The first section reports the descriptive statistics and reliability analysis for the sample. The second section provides a comparison of the original test items written in English, then translated to Japanese. The third section reports the results from the DTF/DIF analysis. The fourth section addresses the possible cause/explanation for DTF/DIF.

Demographic Characteristics of the Sample

The demographic characteristics of the same included 1000 American and 1134 Japanese males and females between the ages of 18 and 40 years. Table 1 provides a descriptive summary for the sample used in this study. Table 2 provides classical item statistics for the Japanese and American samples.

Insert Tables 1 and 2 about here

Differential Test Functioning and Compensatory DIF Results by Item Content

DTF was significant ($\chi^2 = 7015.45$, $p<.0001$) indicating that the two versions of the 30 item test were functioning differentially at the test level. Item number 30 from the skills/rescue content area was found to have the greatest amount of C-DIF (.160). After elimination of item 30, the chi-square statistic for DTF was not significant $\chi^2(1133, N =$

1134) = 1205.97, $p > .05$. Therefore, the only significant item C-DIF item was item 30. Table 3 provides the selected output from the DFIT program.

Insert Table 3 about here

Non Compensatory DIF (NC-DIF) Results

The following items (5 skills/safety, 3 physiology, and 2 physics) were found to have significant NC-DIF values (NC-DIF >.006): 3, 4, 11, 12, 13, 16, 22, 25, 29, 30. NC-DIF assumes that all other items in the test are free of DIF and therefore does not include information about DIF from other items. Therefore, NC-DIF values are particularly good for revealing why certain items exhibit more DIF than others or why various items may be offensive to certain groups.

Comparison of Items 4, 13 and 30 After Translation

Table 4 includes the text for items 4, 13 and 30 in their original English form and the text of those in the Japanese version after translating from Japanese to English word by word.

Insert Table 4 about here

Items 4, 13, and 30 were selected for illustration of DIF detection using the DFIT procedure because these items exhibited the greatest NC-DIF, and item 30 showed significant C-DIF as well.

Judgmental and Empirical Evaluation of DIF

The statistical detection of DIF is only the first step in an analytical process of determining the cause and explanation for such DIF. Significant DIF may be a Type I error, due an artifact of the statistical method, or it may be a sign of multidimensionality related to construct that the test is attempting to measure. In general, after items are flagged for significant DIF, a plausible cause for DIF is speculated by content experts. Finally, if patterns of significant DIF arise in similar item types, the interpretation of DIF is enhanced.

Uniform and Non Uniform DIF

Uniform DIF occurs when the differences in probabilities of success is uniform for the two groups over all ability levels. In this study, items 3, 4, 12, 22, 25, 29 exhibited uniform DIF. Figure 2 illustrates uniform DIF for item 4.

Insert Figure 2 about here

Non-uniform DIF occurs when the probability of success is greater for one group at one end of the ability scale and the probability of success is greater for the other group at the other end of the ability scale. The item characteristic curves for the two groups cross at some point when graphically examined. Figures 3 and 4 illustrate non-uniform DIF for items 13 and 30.

Insert Figures 3 and 4 about here

Items that exhibited non-uniform DIF in this study were 11, 13, 16, and 30. Table 5 identifies how items are classified in relation to both types of DIF.

Insert Table 5 about here

Questions 4, 12, and 25 containing content related to the skills/safety area displayed significant DIF favorable to the Japanese group across the entire ability range. In question 4, the English version of the item stem was written in a negative context, while the Japanese item stem was written in a positive manner (Table 4). Additionally, in item 4, the final word in the stem was spelled incorrectly on the Japanese version. As a result, the DIF detected in item 4 most likely was due to the manner the stem was presented. Questions 3 and 30 related to the skills/safety content area favored the American group consistently at the low end of the ability scale. In question 3, significant DIF appeared even though the stems and possible response choices were precisely the same for both groups (Table 5). Subsequently, there was no logical explanation for DIF in question 3 since both the content and translation for the item were the same. Possibly, the DIF in item 3 was due to a Type I error. In question 13, the semantic nature of both items were the same, but the item formats were different (Table 4). Answer choices "c" and "d" for the Japanese group were

much easier (i.e., obviously false) than the answer choices "c" and "d" for the American group. Therefore, in item 13 different answer choice formats were believed to be a possible cause of DIF.

Item 30 was the only significant C-DIF item in the entire analysis. In question 30, the English version of the stem was written in a positive manner, while the Japanese version of the item stem was written in a negative context (Table 4). Additionally, the answer choices for both groups were different after translation (Table 4). Although the semantic meaning or content of item 30 was intended to be the same for both groups, actual items were quite different in the two versions for this item. A cultural difference may explain a possible cause of DTF. In Japan, the meaning of "one goes first and the other dives afterwards" is interpreted to mean that one person follows directly behind another while walking, swimming, and so on. The sentence implies that one is a leader and the other is a follower. In contrast, the answer choice "c" for the American group reads: "agree on a dive leader" and in America this is interpreted to mean that people dive side by side. For both items the correct answer choice was "c". Finally, for the American group, the item choices were easier to distinguish between the correct and incorrect response than in the Japanese version of item 30.

Question 11, also displaying significant DIF, related to the physiology content area favored the Japanese group at the low end of the ability scale and the American group at the high end of the ability scale (Table 5). In question 11, the translation and semantics of the items and answer choices were the same for both groups. In Question 16, also related

to the physiology content area, the American group is favored at the low end of the ability scale and the Japanese group is favored at the high end of the ability scale (Table 5). The final question in the physiology content area, question 29 displayed DIF favorable to the American group across the entire ability scale. The significant DIF detected in the items related to the physiology content area displayed no logical pattern of favor for either group and not be explained by the translation process or by cultural means.

Question 13 related to the physics content area displayed DIF favorable to the Japanese group at the low end of the ability scale and DIF favorable to the American group at the high end of the ability scale (Table 5). After translation, item 13 yielded the same stem for the two groups, but much easier answer choices "c" and "d" for the Japanese group. In this instance, the translation process could be a cause of the observed DIF. Question 22 from the physics content area displayed DIF consistently favoring the Japanese group over the entire ability scale. In general, the Japanese group performed slightly better on the physics items 8, 13, 15 and 22 (Table 2).

An overall pattern emerges from this information with respect to Japanese students performing better at either all ability levels or at the high end on the questions containing information about general diving skills and safety. On questions 16 and 29, the American group performed better on physiology questions either at the high end of the ability scale or across the entire range of ability. Finally, on question 11, the Japanese group was favored at the high end of the ability scale.

Discussion

The results of this experiment provide significance for several areas of measurement research. First, the results of this project indicate that even translation by a bilingual content expert does not ensure measurement equivalence. For test developers involved in cross-cultural testing situations, the use of the DFIT framework to detect DIF along with judgemental/logical evaluation of item and test DIF/DTF may be helpful. In this project, the DFIT procedure flagged 10 out of 30 items for significant NC-DIF. Thirteen out of 30 of the items displayed translation differences either in the stem and/or in the options, 6 of these 13 items were identified as having significant NC-DIF. This is a much higher rate of DIF (6/13) than the DIF rate for the precisely translated items (4/17). These results suggest that the DFIT procedure effectively identified differential functioning possibly caused by the translation process. The only C-DIF item, which is also a NC-DIF item, was item 30. DTF was no longer significant after removing item 30. What this means to test developers is that if one was to revise this 30 item test, removing or rewriting item 30 would be recommended to keep the test-level differential functioning as small as possible.

Additionally, this study reveals that a logical content analysis of items displaying significant DIF. In this study, logical analysis provided a possible explanation or cause for the observed DIF in most instances. Items displaying DIF resulting from translation errors or cultural insensitivity, may be rewritten and again tested for DIF. If these corrections or modifications eliminate DIF, then the item can be added back to the item pool.

Regarding the issue of test translation protocol, this study provides support for the need to adhere to rigid translation procedures in order to minimize DIF. In this study, only one translator was used to conduct a single translation prior to the test being released for use internationally. The majority of items displaying significant DIF in this study were linked to inaccuracy in the translation process. Therefore, in the future a more precise approach to test translation may reduce the number of items flagged for DIF.

Finally, although some of the post hoc analyses of items displaying DIF are speculative, this study reinforces the idea that possible sources of DIF can be identified using psychometric techniques that are rapidly advancing.

References

Alderman, D.L., & Holland, P.W. (1981). Item performance across native language groups on the Test of English as a Foreign Language (Research report No. 81-16). Princeton, New Jersey: Educational Testing Service.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 1-29). New Jersey: Lawrence Erlbaum Publishers.

Boorstin, D.J. (1985). The Discoverers. New York: Random House.

Bugdell, G. R., Raju, N. S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19 (4) 309-321.

Candell, G. L., & Hulin, C.L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17, pp. 417-440.

Drasgow, F. (1984). Scrutinizing psychological tests: measurement equivalence and equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

Ellis, B.B. (1989). Differential item functioning: implications for test translations. Journal of Applied Psychology, 74 (6), 912-921.

Eysenck, H.J. (1984). The effect of race on human abilities and mental test scores. In C.R. Reynolds & R.T, brown (eds.), Perspectives on Bias in Mental Testing. New York: Plenum, pp. 249-262.

Fleer, P.F. (1993). A Monte Carlo assessment of a new measure of items and test bias. (Doctoral dissertation, Illinois Institute of Technology). Dissertation Abstracts International, 54-04, 2266B.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Hulin, C.L. (1987). A psychometric theory of evaluations of item and scale translations. Journal of Applied Psychology, 67.

Hulin, C.L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18. pp. 115-142.

Hulin, C.L., & Mayer, L.J. (1986). Psychometric equivalence of a translation of the job description index into Hebrew. Journal of Applied Psychology, 71 pp. 83-94.

Jensen, A.R. (1980). Bias in Mental Testing. New York: Free Press.

Lee, K., & Oshima, T.C. (1996). IPLINK. (computer program). Atlanta: Georgia State University.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias is a test of reading comprehension (Technical Report No. 163). Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Lord, F. (1980). Applications of Item Response Theory to Practical testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mercer, J.R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In C.R. Reynolds & R.T. Brown (eds.) Perspectives on bias in mental testing. New York: Plenum, pp. 293-256.

Messick, S.A. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement, (3rd ed., pp. 13-103). New York: MacMillian.

Raju, N. (1988). The area between two item characteristic curves. Psychometrika. 53 (4), 495-502.

Raju, NS., van der Linden, W.J., & Fleer, P.F. (1992). An IRT-based internal measure of test bias with applications for differential item functioning. Paper presentation at the American Educational Research Association, San Francisco.

Raju, N.S., van der Linden, W.J., & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement. 19 (4), 353-368.

Shepard, L.A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & C. Modgil (Eds.), Arther Jensen: Concensus and controversy (pp. 170-190). New York: Falmer Press.

Stout, W. (1991). DIMTEST (computer program). Champaign: University of Illinois.

van de Vijver, F. & Poortinga, Y. (1991). Testing across cultures. In R.K. Hambleton & H. Wainer (eds.) Advances in educational and psychological testing. Boston: Kluwer Academic Publishers, pp. 277-304

Table 1

Descriptive Statistics and Reliability for Sample

Group/Gender	n	Mean	SD	Reliability
American	1000	26.7	1.9	.37
Male	422	26.7	2.0	
Female	631	26.7	1.7	
Japanese	1134	27.5	1.9	.47
Male	631	27.0	1.9	
Female	503	26.8	2.0	

Note. Coefficient Alpha was used in the computation of the reliability index.

Table 2

Summary of Classical Item Analysis for Japanese and American Samples

	Japanese		American	
Item	Percent Correct	Biserial Correlation	Percent Correct	Biserial Correlation
1	.990	.285	.968	.010
2	.991	.210	.999	-.153
3	.896	.323	.994	.388
4	.893	.060	.722	.183
5	.990	.086	.986	.235
6	.974	.102	.949	.253
7	.987	.001	.969	.186
8	.981	.347	.979	-.025
9	.976	.037	.971	-.056
10	.938	.321	.961	.240
11	.902	.214	.883	.409
12	.811	.139	.729	.077
13	.674	-.055	.557	.202
14	.913	.030	.913	.193
15	.908	.444	.948	.173
16	.833	.318	.928	.173
17	.926	.260	.927	.013
18	.930	.357	.958	.346
19	.932	.147	.903	.142
20	.945	.304	.945	-.139
21	.843	.248	.795	.065
22	.824	.111	.678	.144
23	.906	.174	.827	.115
24	.966	.248	.918	.090
25	.969	.135	.818	.194
26	.935	.324	.907	.019
27	.915	.130	.846	.065
28	.938	.237	.994	.624
29	.887	.416	.974	.000
30	.856	.173	.917	.012

Table 3

Differential Functioning of Items and Tests

Number of Examinees:	1134	
Item	C-DIF	NC-DIF
1	-.006	.000
2	.002	.000
3	.022	.017
4	-.041	.025
5	.001	.000
6	-.001	.001
7	-.002	.000
8	-.000	.000
9	-.000	.000
10	.009	.001
11	.015	.008
12	-.008	.008
13	.022	.036
14	.010	.002
15	.009	.004
16	.037	.012
17	.004	.000
18	.010	.001
19	-.002	.001
20	-.005	.001
21	-.008	.001
22	-.028	.019
23	-.014	.006
24	-.012	.002
25	-.039	.019
26	-.009	.001
27	-.016	.003
28	.016	.004
29	.024	.012
30	.160	.271

(table continues)

	True-F	True-R	D	C-DIF
Mean	27.074	27.429	-.35425	.00499
Variance	2.057	2.225	.02420	.00111
SD	1.434	1.492	.15555	.03337
Differential Test Functioning	.14969			
Chi-Square	7015.45			
Probability	.0000			
Degrees/Freedom	1134			

Table 4

Items 4, 13 and 30 Translated From English to Japanese

Item	Language	Question Stem
4	English	The least desirable dependent option in an out-of-air situation is buddy breathing. a) T b) F
4	Japanese	During the emergency procedure for out of air, the most recommended method for getting assistance coming up is to get the optional second slage. a) T b) F
13	English	To maintain neutral buoyancy during descent, a diver wearing a wet suit should: a) Add air to the BC. b) Dump all the air from the BC. c) Activate the J-valve. d) Remove some lead from the weight belt.
13	Japanese	During descent, a diver wearing a wet suit should _____ to maintain neutral buoyancy. a) Add air to the BC b) Dump all the air from the BC c) Hold a rock instead of a weight d) Get rid of all of the air in the lungs and hold your breath
30	English	It is good practice for diving buddies to: a) Wear matching equipment. b) Have the same certification level. c) Agree on a dive leader. d) Practice emergency swimming ascents.
30	Japanese	Which is wrong concerning the buddy system? a) Go down and up together always. b) Swim side by side in the distance that you can reach each other by hand. c) In the water, one goes first and another dives afterwards. d) Decide beforehand which one will take the leadership.

Table 5

DIF Identification and Classification

Item/Content	DIF	Ability Scale Location	Group Favor	Translation Result
3- skills/safety	uniform	low	American	no difference
4- skills/safety	uniform	entire range	Japanese	mispelling/ negative vs. positive stem
11- physiology	non-uniform	low high	Japanese American	no difference
12- skills/safety	uniform	entire range	Japanese	different answer choices
13- physics	non-uniform	low high	Japanese American	different answer choices
16- physiology	non-uniform	low high	American Japanese	no difference
22- physics	uniform	entire range	Japanese	different answer choices
25- skills/safety	uniform	entire range	Japanese	different stem/ different answer choices
29- physiology	uniform	entire range	American	no difference
30- skills/safety	non-uniform	low high	American Japanese	negative vs. positive stem/ different answer choices

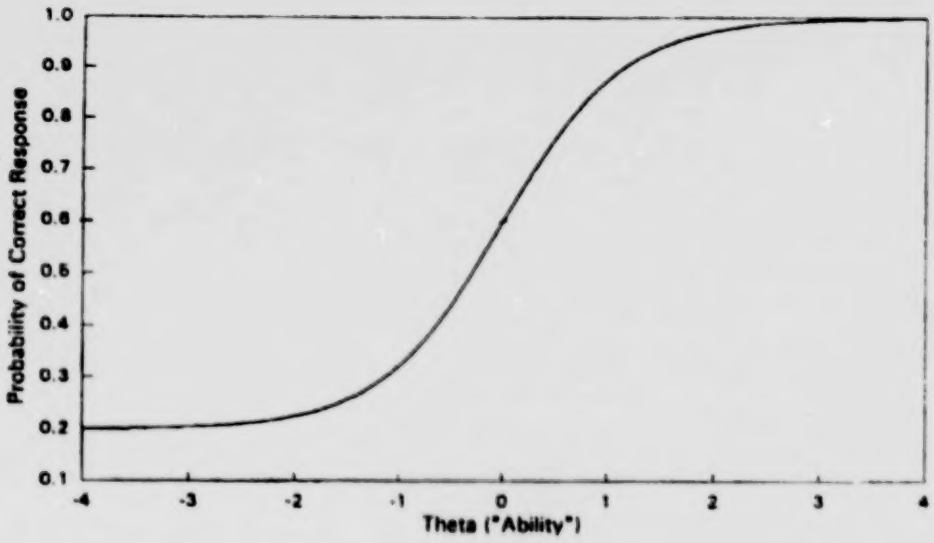

Figure 1. Item characteristic function with a = .425, b = 0, and c = .20.

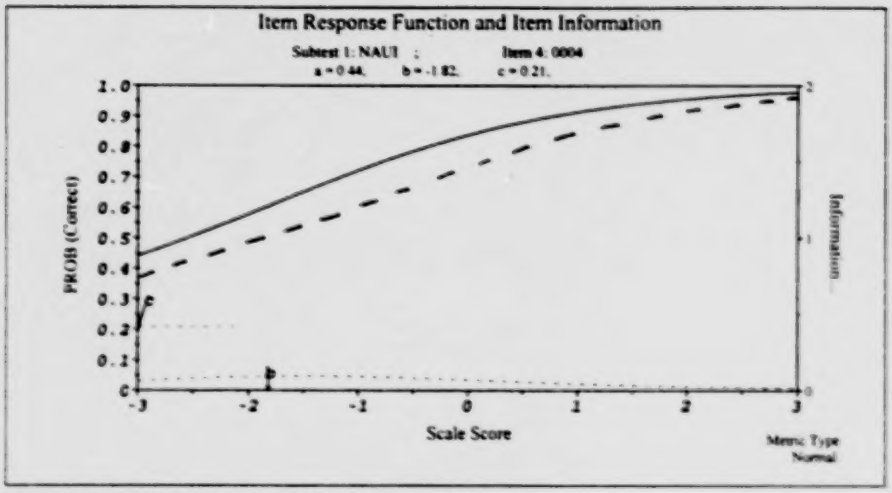

Figure 2. Item characteristic curves for item 4.

American - - - - - - a= .42 b= -.83 c= .25

Japanese ——— a= .44 b= -1.82 c= .21

*Information curve is the dashed line across the bottom.

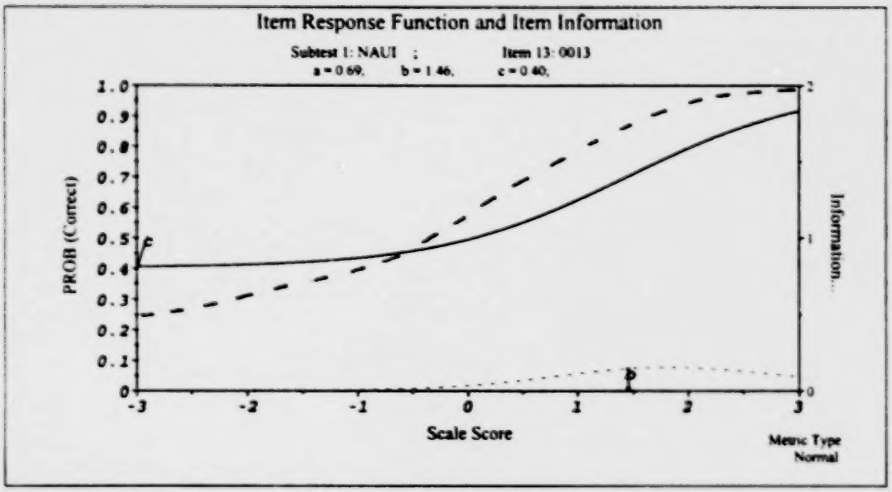

Figure 3. Item characteristic curves for item 13.

American - - - - - - a= .67 b= .20 c= .19

Japanese ——— a=.69 b= -1.46 c= .40

*Information curve is the dashed line across the bottom.